基于 BERTopic 的突发事件微博舆情主题演化分析研究——以“东航 Mu5735 空难事件”为例*

江长斌，徐思思，黄英辉，徐紫琪，王宏宇
(武汉理工大学 管理学院，湖北武汉 430070)

摘要：[目的/意义]本研究旨在系统地分析突发事件微博舆情的主题演化趋势，可视化舆情发展过程中的焦点主题，为后续网络舆情的引导提供实践参考依据。**[方法/过程]**采用 BERTopic 主题提取模型识别舆情发展不同阶段的主题，并采用余弦相似度度量主题间的相似性，可视化主题的演化路径。以新浪微博“东航 Mu5735 空难事件”为例，进行突发事件网络舆情的主题演化研究。**[结果/结论]**实证研究结果表明，基于 BERTopic 主题模型对舆情事件建模得到高效且可观性较高的主题识别结果，准确把握舆情发展每个阶段中热点主题，揭示了舆情传播过程中主题演变过程。**[创新/局限]**本研究提出了一种基于 BERTopic 模型对短文本突发事件微博舆情主题演化分析的总体框架，对主题提取后的结果进行主题内容演化分析并进行可视化展示。本研究的局限性在于当前选用的数据来源仅来源于微博平台，后续可提高数据来源的多样性。

关键词：突发事件舆情；BERTopic 主题模型；主题识别；主题演化
分类号：G206

1 引言 / Introduction

在网络信息快速交互发展的时代，微博等社交媒体已成为信息传播的重要平台。微博因其快速、开放和互动性强的特点，使其成为突发事件爆发下舆论生成的主要媒介，各类信息被广泛地传播和讨论。在突发公共事件中，微博平台不仅能够迅速传播事件信息，还能够反映甚至放大公众的情绪和意见，对突发事件舆情的治理产生重要影响。但迅猛的信息传播速度和广泛的用户参与度，带来了信息准确性传达、事件谣言控制和公众情感引导等方面的挑战。基于此，探究微博平台中突发公共事件网络舆情的主题演化过程，挖掘该类事件背景下不同生命周期阶段舆论的关注焦点，为政府和相关机构提供科学的决策依据，更有效地应对和管理突发事件中的网络舆情，促进信息的健康传播，维护网络空间的秩序。本研究选取重大突发公共事故“东航 Mu5735 空难事件”为案例，采用 BERTopic 主题提取模型对微博平台中相关评论数据进行主题挖掘和事件分析，并提供可视化的数据分析结果，把控突发公共事件舆情发展方向、正确引导网络舆论、提供可供参考的相关部门治理策略。

---

*本文为国家社会科学基金项目“大数据视域下‘隐性’政治舆情演化规律及治理路径研究”的研究成果之一，项目编号：19BSH013。
作者简介：江长斌，副教授，硕士生导师；徐思思，硕士研究生；黄英辉，副教授，博士；徐紫琪，硕士研究生；王宏宇，副研究员，博士，通讯作者，E-mail：hongyuwang@whut.edu.cn。

2 文献综述 / Literature review

2.1 突发事件网络舆情研究

突发事件网络舆情的演化具有突发性、难以预测性、关注度高、扩散性强、影响力大等特点，为更好地应对突发事件网络舆情，需要对突发事件网络舆情的演化模式和规律进行深入研究[1]。当前国内外对于突发事件网络舆情的研究主要集中于以下三个方面：一是对于舆情中话题的发现及其演化分析。Zhong[2]采用 LDA 主题提取分析、SnowNLP 公众情感分析和相关性分析，对新冠肺炎突发公共卫生事件进行舆情演化研究。Chen 等[3]基于 Adam-LSTM 模型预测突发公共事件中网络舆情的热度演化。许露萌[4]以时空视角运用全局空间自相关分析、局部空间自相关分析和灰色关联分析等，进行突发公共卫生事件网络舆情的热度演变分析。王晰巍等[5]利用事理图谱对重大突发事件舆情下 UGC 进行分析，实现舆情事件之间的因果演化过程及演化路径可视化。张柳等[6]基于 LDA 模型对国内外突发事件应急管理文献进行主题挖掘，分析主题热度及确定主题演化路径。曹树金等[7]结合生命周期理论、TF-IDF 特征词-权值模型以及潜在狄利克雷模型方法，探索突发公共卫生事件微博舆情传播周期中各阶段的热点主题，勾勒舆情事件主题演化的时序发展趋势。二是对于突发事件中网络舆情的触发机制和原理的研究。周林兴等[8]利用事理图谱有效解析了重大突发事件网络舆情诱发与缓释机理。杨洋洋[9]基于社会压力—事件状态—舆情响应（PSR）的视角，探究了舆情事件中原因要素对重大突发事件中网络舆情触发的影响机制。杨洋洋等[10]采用 NCA 与 fsQCA 相结合的融知发酵模型，探究突发灾难事件舆情的发酵机理和理论模式。三是对于突发事件下网络舆情分析中群体极化现象的研究。卢国强等[11]以新冠疫情期间“辉瑞新冠小分子药物”事件为例，构建了极端观点的 TCMCR 识别模型，对其网络群体极化中极端观点进行了有效识别。卢国强等[12]基于风险耦合的角度分析三元空间信息观领域下突发事件网络舆情群体极化的形成因素，并对其进行了群体极化的演化分析。贾若男等[13]根据“数据—知识—服务”的转化路径构建了突发事件网络舆情的群体极化风险评估模型。

2.2 主题建模技术研究

主题建模作为一种高效的文本分析工具，能够从大规模的网络文本中提取出潜在的主题，为网络舆情分析提供了深入的见解。目前国内外对于文本主题建模分析的研究方法包括从传统算法到最新的基于深度学习的技术。一是基于文档—单词的共现频率特征来抽取主题的词袋模型，如潜在语义分析[14]（LSA）、概率潜在语义分析[15]（pLSA） 和隐含狄利克雷分布[16]（LDA）等。田世海等[17]采用改进潜在语义分析和支持向量机算法（LSA+SVM），构建用于突发安全

事件舆情分类的预警模型。周楠等[18]结合 PLSA with Background Language 与关键词聚类的方法发现舆情事件内部子话题，并最终生成事件子话题标签的 ET—TAG 模型，挖掘事件共性以及反映事件子话题热度的变化趋势。曾子明等[19]利用 LDA 模型探究突发公共卫生事件网络舆情各周期间的舆情主题差异。Yu 等[20]应用 LDA 模型对研究中所选取的人工智能领域文献的摘要进行主题提取，探究人工智能领域的主题和趋势。Sakshi 等[21]使用 LDA 对 1967 年—2021 年发表的 325 篇研究论文的语料库进行了主题提取，从而确定数学表达式识别的最新发展趋势。二是基于预训练词嵌入的聚类方法，从嵌入的聚类空间的簇中去采样主题词，如 Top2Vec[22]模型和 BERTopic 模型。Ghasiya 等[23]采用 Top2Vec 模型对四个国家有关于新冠肺炎的新闻报道进行主题提取，挖掘关键主题并探索演化趋势。Grootendorst[24]提出了 BERTopic 主题建模方法，解决传统算法中的局限性：采用 BERT 句子转换器制造高质量、包含上下文语义的句子矢量表示；采用基于类的 TF-IDF 弥补聚类与采词空间前后不兼容问题。Wang 等[25]提出了基于 BERTopic 的跨学科主题识别和演化分析的框架，实现从微观层面对跨学科主题演变过程的研究。曹树金等[26]利用 BERTopic 模型对所选取的信息资源管理学科四个数据集中文献的摘要进行主题识别，探究信息资源管理学科发展方向。

综上所述，对于突发事件网络舆情的研究仍是当下的热点之一，尤其是涉及演化过程的研究，对于该类舆情的初步处理是需进行主题识别以及演化路径展示。当前舆情领域使用最为广泛主题识别的方法是 LDA 模型，其通过使用狄利克雷先验概率改进了 pLSA，为新文档分配一个概率，克服了 LSA 和 pLSA 的局限性，但 LDA 仍具有传统主题提取方法的典型问题且对于主题的数量需进行提前设定。Top2vec 模型和 BERTopic 模型分别使用 Doc2Vec 方法、BERT 句子转换器来实现从输入文档中制造语义高质量嵌入，且针对社交媒体帖子以及评论这类短文本的识别效果更优。BERTopic 模型在进行主题识别时涉及的步骤是自洽的，可根据该领域的进展以及特定的项目或技术限制进行自主选择；其支持分层减少主题以优化主题的数量，即主题的数量不一定事先给定；提供先进的内置搜索和可视化功能，为演示制作了高质量多种类的图表。通过有效地提取和分析大量文本数据中的主题，主题识别模型技术揭示公众关注的焦点，预测舆情发展趋势，理解和监测公众对突发事件的反应和情绪，为舆情管理和危机应对提供理论支持。

3 研究设计 / Methodology

3.1 研究思路

本研究提出了一种基于 BERTopic 模型对短文本突发事件微博舆情主题演

化分析的总体思路。以某一微博舆情突发公共事件为研究对象，选取相关话题下用户评论文本为数据源，探究突发公共事件网络舆情的主题演化趋势。研究思路主要分为以下四个部分：数据收集、数据预处理和事件阶段划分、主题模型识别、主题演变分析，如图 1 所示。

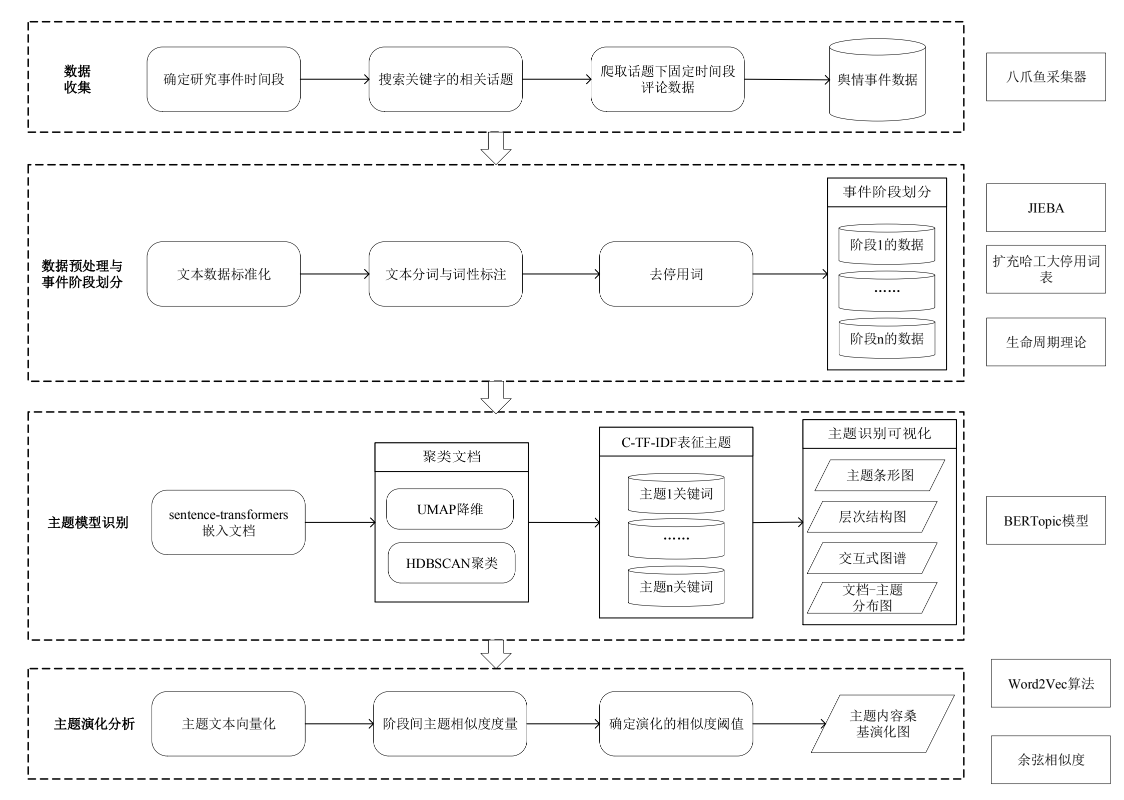


图 1 主题演化分析总体思路图

Figure 1 Research idea of thematic evolution analysis

数据收集：在微博上搜索某事件关键字的相关话题，确定研究事件时间段，采用八爪鱼采集器收集评论数据。

数据预处理与事件阶段划分：对爬取到的数据进行数据预处理，并且结合生命周期理论和百度指数趋势图进行舆情事件阶段划分。

主题识别与演化：采用 BERTopic 模型对处理后的数据进行主题提取以及可视化；对提取到的主题文本进行 Word2Vec 向量化，计算主题相似度，判定主题间的演化关系并可视化。

3.2 研究方法

3.2.1 BERTopic 主题提取模型

BERTopic 模型克服了传统主题识别模型的两种局限性：未考虑单词的上下文语义；聚类和采词之间不一致。BERTopic 模型采取预训练的语言模型创建文档嵌入从而获取文档级别的信息，充分考虑上下文语义；使用一种基于类别的 TF-IDF 变体进行主题表征，解决传统视角下聚类和采词之间的不兼容问题[24]。

BERTopic 模型进行主题识别尤其是针对短文本数据主要经过以下三个模块化的步骤，以下具体介绍每个步骤的流程。

（1） 嵌入文档

将文档转换为数字进行表示。BERTopic 模型中默认使用 sentence-transformers 模型，该模型使用预训练的语言模型将句子和段落转换为密集的向量表示，并且针对语义相似性进行一定程度的优化，便于后续的聚类。在嵌入文档的过程中需选择适合文本对应的语言模型，常用的有以下两种：英文以及其他多语言模型。其中针对英文文档选定的是英文模型"all-MiniLM-L6-v2"；本研究的文本数据都是中文，因此选定的是多语言模型"paraphrase-multilingual-MiniLM-L12-v2"的设置。

（2） 聚类文档

对所得的词嵌入向量进行文档聚类。首先，在进行聚类之前采取降维处理，解决维数过高导致嵌入空间稀疏，聚类模型难以聚类的问题。在 BERTopic 模型中默认设置是 UMAP 算法， UMAP 经证明可以在较低的投影维数中保留更多的高维数据的局部和全局特征结构[27]，从而保留创建语义相似文档集群所需的信息。与此同时，UMAP 算法对嵌入维度没有计算限制，适用于不同维度空间的语言模型。

其次，采用基于层次和密度的 HDBSCAN 算法对降维后的向量进行分类，从而得到各个主题簇。HDBSCAN 使用软聚类方法建模集群，具有在可能的情况下识别异常值的功能，允许将噪声建模为离群值，从而防止不相关的文档分配给不适用的集群，提取的噪声更少，聚类质量更高。

（3） 表征主题

采用基于类的 TF-IDF 变体（c -TF -IDF）算法提取聚类后的每个簇的主题词，挖掘重要词汇，从而实现主题的表征。使用 HDBSCAN 作为聚类模型后，所得的集群是具有不同程度的密度和结构， BERTopic 模型中采用 c -TF -IDF 算法，它对集群的预期结构没有任何要求。

为了将采词空间收束到对应的簇上，首先是将集群中的所有文档合并为一个长文档，该长文档表示群集；然后，计算每个单词在每个集群中出现的频率；最后，通过 TF 与 IDF 的频率相乘得到词汇的重要性得分，合并 top c-TF-IDF 中相似的主题表示。通过以上流程便得到 c -TF -IDF 算法，其计算公式如公式（1）所示。

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t}) \quad (1)$$

其中，c 表示 class，A 表示每个 class 的平均单词数量，class c 是为每个集群连接成单个文档的文档集合，$tf_{t,c}$ 表示 class c 中词 t 的频率，

tf_t 表示所有 class 中词 t 的频率。

3.2.2 文本向量化与相似度度量

BERTopic 模型对三个阶段的文本分别进行主题提取，通过 c -TF -IDF 算法获得每个主题对应的重要主题词。首先对提取到的主题词利用 Word2vec 算法进行向量化表示，再采用余弦相似度计算相邻两阶段的主题相似性，确定主题间可进行演化的相似度阈值，得到演化路径，进而揭示舆情演化的规律，为后续舆情的引导策略制定提供理论依据。

4 实证研究 / Empirical research

4.1 数据收集

本研究在微博上以“东航 Mu5735”为关键字获取相关话题，整个事件的采集的时间段选取是基于表 1 中时间线以及事件搜索的百度指数，即从 2022 年 3 月 21 日事故发生到 2022 年 4 月 20 日“3 • 21”东航 MU5735 飞行事故调查初步报告发布后一天 4 月 21 日为固定时间段。通过八爪鱼采集器爬取每个话题下固定时间段的全部评论数据，包括评论人、评论内容以及评论时间。

表 1 “东航 Mu5735”事件梳理时间线

Table 1 Timeline of the "China Eastern Airlines Mu5735" Incident

时间	事件
2022-03-21	一架客机在广西藤县**发生事故**，并引发山火，伤亡情况未明。
2022-03-21	民航局**确认**东航一架**飞机坠毁**；机上人员共 132 人，其中旅客 123 人、机组 9 人。
2022-03-22	截至 3 月 21 日 24 时，**尚未发现幸存者**，飞机的**黑匣子**也**尚未找到**。
2022-03-23	新闻发布会上通报，东方航空公司 MU5735 航班的**一部黑匣子**已于 23 日被**发现**。
2022-03-23	东航坠机事故，消防救援人员已**发现部分飞机残骸和人体组织碎片**。
2022-03-26	东方航空公司 MU5735 航班上人员**已全部遇难**。
2022-03-27	失事飞机**第二部黑匣子已找到**。
2022-04-20	民航局发布关于“3•21”东航 MU5735 飞行事故**调查初步报告**的情况通报。

4.2 数据预处理与事件阶段划分

本研究对文本数据的预处理主要是进行分词和去停用词处理、对应的日期进行格式统一化。首先对于文本数据标准化，包括进行繁体简化、字符整数型转换为字符串型、异常字符与格式过滤、表情符号与转发评论标识正则表示化等。其次对文本数据内容借助 JIEBA 分词工具进行分词与词性标注，最终筛选保留两个字及以上的名词、人名、地名、机构名、其他专名、形容词、动词和名动词的结果。最后采用哈工大停用词表并进行扩充，对分词后的结果进行去除停用词处理，经过以上预处理后，获得 10120 条文本评论数据。

从“东航 Mu5735”事件的百度指数关键词搜索趋势图（图 2）观察得出：“东航 Mu5735”的活跃期以 2022 年 3 月 21 日为起点、以 2022 年 4 月 21 日为终结点，且其舆情传播过程呈现出明显的三个阶段的变化。本研究依据生命周期理论和百度指数趋势图，将整个舆情事件划分为以下三个阶段：发生期（3 月 21 日—3 月 23 日）、扩散期（3 月 24 日—3 月 27 日）、消退期（3 月 28 日—4

月 21 日）。

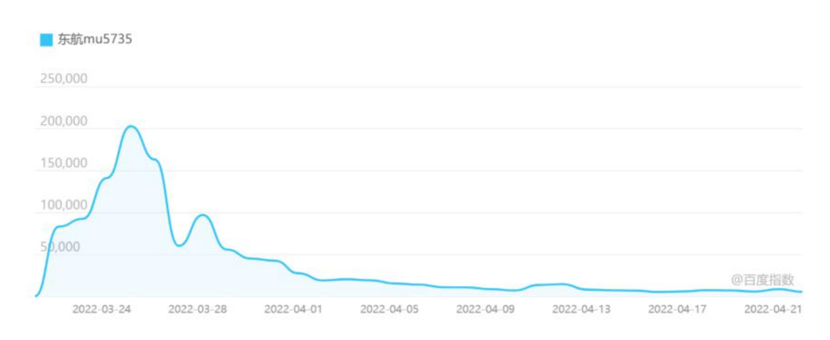


图 2 “东航 Mu5735”舆情传播趋势图

Figure 2 Trend of " Eastern Airlines Mu5735" Public Opinion Dissemination Chart

4.3 主题提取

4.3.1 发生期主题提取

在发生期期间主要是事故发生引发山火，消防队以及应急人员开展救援，调查组以及相关媒体进行跟踪报道现场搜救情况，直至 23 日确认所有机上人员全部遇难。对发生期间的文本数据进行主题提取，共获得 17 个主题，通过 c-TF-IDF 计算得到每个主题对应的特征词，可视化结果采用条形图展示前 16 个主题的主题特征词，如图 3 所示。其中提取到的各主题词分类主要分为表 2 所示，在突发事故发生初期网民集中关注于事故发生过程、搜救情况、失联者家属心理疗愈等。

表 2 发生期主题词类别

Table 2 Occurrence Period Category of Thematic Feature Words

主题	方向	主题	方向
Top0	网民祈祷平安	Top8	机上人数确认向
Top1	事故定性向	Top9	幸存者生还
Top2 与 Top12	谣言猜测向	Top10	家属心理疗愈向
Top3	跟进报道向	Top11	坠机地点起火向
Top4	失联者家属安抚	Top13	现场搜救直播向
Top5	地域身份向	Top14	搜救过程
Top6	搜救现场环境	Top15	机型信息
Top7	黑匣子搜寻向	Top16	失联者随身物品

借助图 4 可以直观挖掘主题间的潜在联系，进行主题的合并。主题 2 与主题 12 都是围绕着舆情发生期的期间相关谣言猜测向；主题 11、主题 13 与主题 16 都是围绕事故现场救援过程展开；主题 8、主题 9 与主题 6 之间的联系可以解释为，基于搜救现场的现状而引起对机上人员总数猜测；主题 4 与主题 10 是对失联者家属救助以及安置的问题上潜在的联系；主题 0、主题 14 与主题 5 之间的联系为网民密切关注搜救过程从而表达祈求平安以及搜救现场出现失联者地域身份信息从而引起同地域网友的共情。针对主题 1、3、7、15 主题之间的联系可以理解为：面对突发的空难事故，东航、民航局等首先要核实事故飞机的基本航班信息，其次对产生的舆情风波召开发布会，及时做出回应；网民最为迫切关怀的仍有事故真相，而破解真相的关键在于黑匣子的破译，因此与主

题 7 黑匣子搜寻有着间接的关系。黑匣子的破解技术与机型信息（Top15）有着紧密关系，东航 MU5735 是波音 737 系列客机。

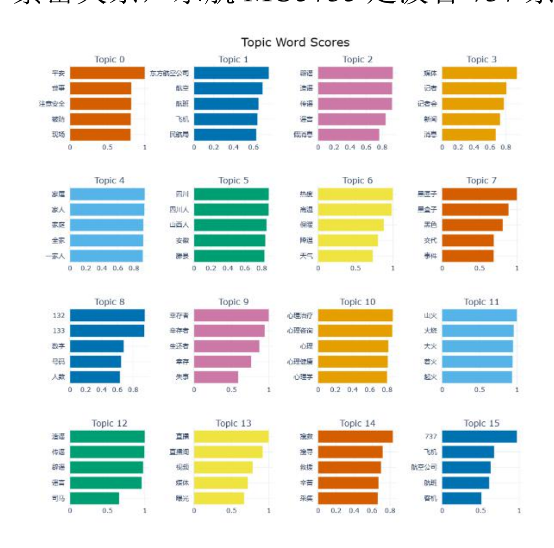


图 3 发生期的主题特征词分布（前 16 个）

Figure 3 Distribution of Thematic Feature Words for the Period of Occurrence （Top 16）

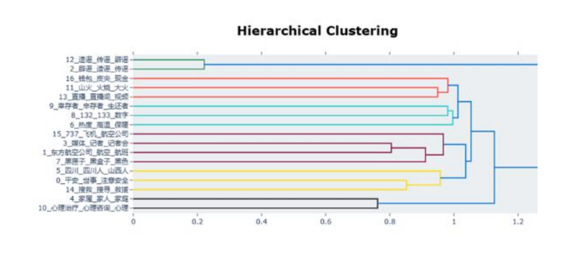


图 4 发生期的主题潜在的层次结构

Figure 4 Potential Hierarchy of Themes in the Period of Occurrence

4.3.2 扩散期主题提取

由于前期的搜救结果并未发现任何失联者的遗体以及黑匣子，但是在 3 月 23 日接连发现第一部黑匣子以及人体组织碎片，使得事故调查进度和搜救结果出现了实质性进展，将事故舆情发展推往高潮。整个期间的文本评论内容进行主题提取，最终共获得以下图 5 展示的 15 个主题，对主题词的分类如表 3 所示。该阶段主题可以总结为以下四个方面：一是经过 DNA 比对，确认机上人员全部遇难，引发网民对遇难者的缅怀和祭奠，并更加迫切转发关注事故真相；二是针对事故发生后，中国面临的外部舆论环境复杂，别有用心者借以美国阴谋论来挑拨中美关系；三是关注黑匣子的破译进度，包括黑匣子严重受损极大阻碍了破译进度使得网民发布大多情绪否定词的评论；四是对于遇难者信息公布与否，从而引发尊重家属的意愿和保护遇难者的隐私两个方向的言论。

表 3 扩散期主题词类别

Table 3 Diffusion Period Category of Thematic Feature Words

主题	方向	主题	方向
Top0	事件转发关注	Top8	黑匣子破译
Top1	航空信息	Top9	遇难者祭奠
Top2	失联者已逝	Top10	情绪否定词
Top3	中国空难事故	Top11	黑匣子受损
Top4	美国阴谋论	Top12	航班信息
Top5	媒体跟进报道	Top13	遇难者信息公开意愿
Top6	官方信息公布	Top14	事故真相
Top7	遇难者家属救助		

对扩散期的主题进行可视化描述，生成交互式图谱如图 6 所示。其中每一个圆圈都代表一个主题，圆圈的大小表示主题在所有文档中出现的概率。图中

事件转发关注（Top0）、航空信息（Top1）、失联者已逝（Top2 ）三个主题在文档中出现的概率较高的主题；圆圈之间距离的远近表示主题之间相似性的程度，如遇难者家属救助（Top7）与遇难者信息公开意愿（Top13）两个圆圈距离较近其对应的相似程度较高，余弦相似度求解相似度值为 0.82，两者都是基于遇难者家属对接处理方面的做法。

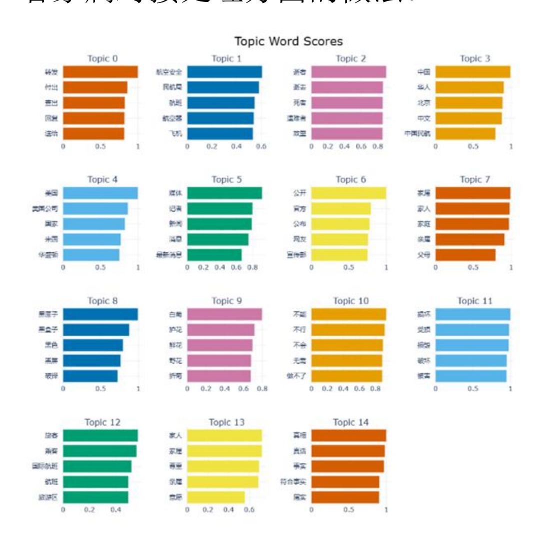


图 5 扩散期的主题特征词分布

Figure 5 Distribution of Thematic Feature Words in the Diffusion Period

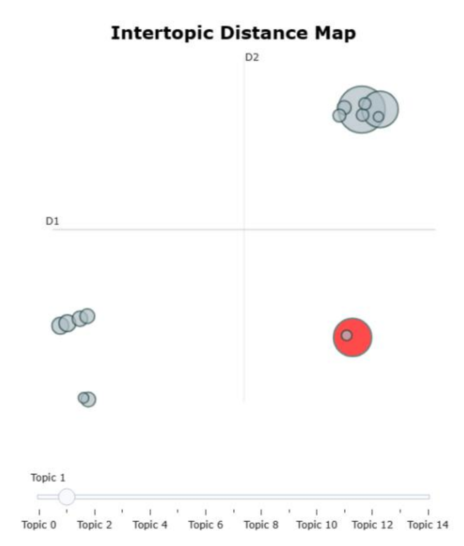


图 6 扩散期的主题交互式图谱

Figure 6 Interactive Mapping of Themes in the Diffusion Period

4.3.3 消退期主题提取

随着时间的推移加之空难事故调查持续时间长，整个研究事故时间线截至于 4 月 20 日民航局公布事故调查初步报告后一天时间，舆情逐渐淡出公众视线结束。该期间进行主题提取后，共获得以下 17 个主题，前 16 个主题的主题特征词分布如图 7 所示，对主题词的分类如表 4 所示。该阶段的主题聚焦于事故调查以及遇难者缅怀两个方面。首先，民航局公布事故调查的初步报告，对事故领域定性、坠机轨迹描述、黑匣子破译进度以及针对初步报告后对信息的有效性与可读性的评判。其次，消退期间正经清明时节，无情的空难使得这个清明节更加伤感与痛心，网民纷纷以各种方式表达对遇难者的哀悼对家属的安慰，包括发表让家属借以时间慢慢释怀疗愈悲痛的看法；以及在进行事故调查时遇难者个人隐私信息公开征求家属意见相关评论。

表 4 消退期主题词类别

Table 4 Fading Period Category of Thematic Feature Words

主题	方向	主题	方向
Top0	事件转发关注	Top9	事故真相
Top1	航空信息	Top10	遇难者隐私信息保护
Top2	调查报告通报	Top11	事故调查进度
Top3	情绪否定词	Top12	事故原因
Top4	遇难者缅怀	Top13	间疗愈法
Top5	黑匣子破译难点	Top14	等待事故真相

Top6	坠机事故	Top15	坠机轨迹描述
Top7	遇难者哀悼	Top16	事故报告评判
Top8	清明祭东航		

为了快速识别并可视化消退期中文本数据集中的讨论主题的分布情况，生成图 8 所示的文档-主题分布图谱，每个点代表一个文档，同种颜色的簇是代表同一个主题下的文档，颜色不同代表的主题不同。根据图谱，不同颜色的点聚集是表明这些主题有着较高的相似性，同时也是最引人关注的讨论点，如事故真相、调查结果、事故调查进展以及航空安全措施这些主题的集中讨论。作为最为关注的主题群，是由于民航局针对该事故公布了初步调查报告，但黑匣子受损严重破译难度较大，事故真相短时间无定论，因此针对该调查结果会有颇多讨论，网民表达一些包括负面情绪评论以及调查报告结果的认可程度。



图 7 消退期的主题特征词分布

Figure 7 Distribution of Thematic Feature Words in the Fading Period

图 8 消退期的文档-主题分布图谱

Figure 8 Document-Topic Distribution Mapping for Fading Period

4.4 主题演化

通过 BERTopic 模型进行主题建模获得三个阶段各自对应的主题分布信息，再对相邻阶段的主题采取余弦相似度进行主题间相似性度量，设定舆情演化的相似度阈值，采用桑基图揭示整个事件舆情发展过程主题演化路径。其中阈值的设定是结合多次实验验证以及结合实际情况下主题演化规律，最终选定了 0.2。整个过程中主题演化的结果如图 9 所示，图中线条的宽度代表主题间余弦相似度的大小，直观的流线图展示了不同阶段舆情话题的演变与转换，助于追踪和理解舆论焦点的变迁，对于预测舆情走向和制定治理策略提供了理论参考意义。

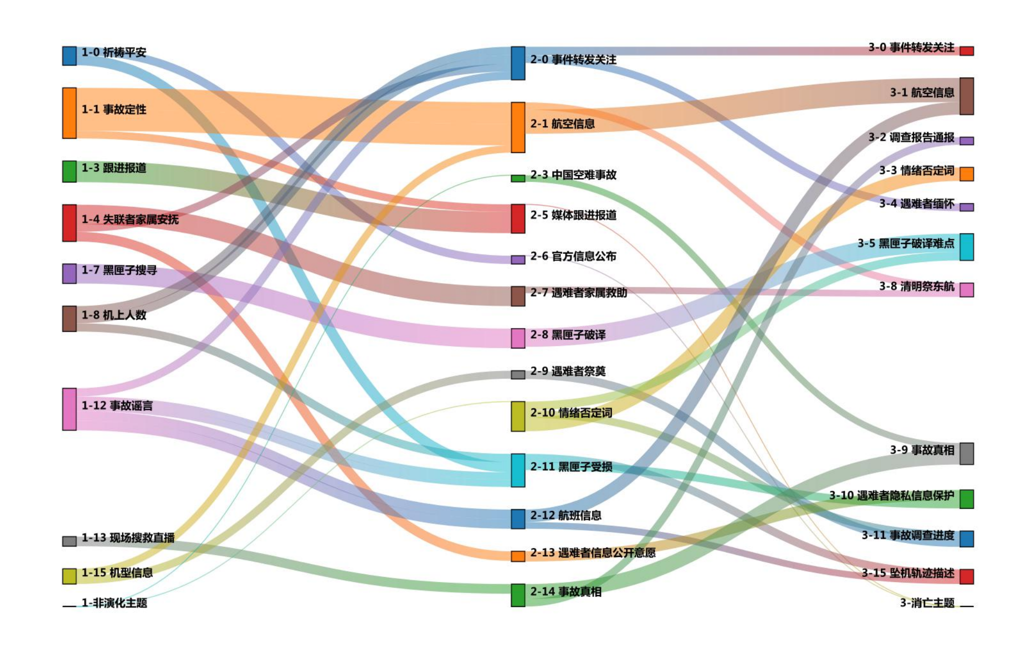


图 9 主题内容演化桑基图

Figure 9 Thematic Content Evolution Sankey Chart

从发生期演化到扩散期的主题主要是有关事故基本情况（1-1、1-8 与 1-15）、救援过程关注（1-0、1-3、1-7 与 1-13）、失事乘客家属安抚（1-4）以及事故谣言（1-12）四类方向。对事故基本情况和搜救过程关注演化至搜救结果公布，如官方和媒体公布搜救结果信息、确认失联者皆已遇难以及黑匣子现状；失事乘客家属转发关注后续救助理赔工作以及征求家属对于遇难者信息公开意愿；对事故谣言方向演化至直击谣言探究真相等事件如黑匣子和航空领域信息。

扩散期中出现了新生演化主题 3（中国空难事故）和主题 10（情绪否定词）。主题 3 的出现是基于在扩散期间事件发酵到引起了国外的关注，对外部舆论以中国空难事故做出回应。主题 10 中大量的情绪否定词是在这阶段搜救结果出现了实质性进展，但所有失联者全部遇难以及黑匣子受损严重，搜救结果的强烈悲剧性，形成了一股强大的情绪否定意义的舆论悲情力量。主题 5（媒体跟进报道）与主题 6（官方信息公布）在消退期演化过程中消亡了。针对主题 5 媒体主要是对搜救过程进行全程关注报道，在该阶段已经得到了明确的搜救结果，搜救过程便告以段落，主题演化结束；针对主题 6 中舆情信息要以官方信息公布为准的消亡原因是，空难事故调查本身复杂性以及解读具有较高的专业壁垒，民间舆论非理性发言得到一定的抵制，呼吁一切以官方公布信息为主，纠正舆论偏差。

消退期的主题大多演化到事故调查初步报告通报、清明时节悼念遇难者两个方向。该事故调查初步报告是整个消退期的舆情讨论焦点，报告通报事故发生的基本航空信息、坠机过程以及事故现场场景、事关真相的黑匣子记录器的

破译进度等，但对于事实真相仍无定论网民发表带有否定情绪意义的评论。对于遇难者家属对接处理是始终贯穿全阶段，以及清明时节全网沉痛情绪表达来悼念此次事故的遇难者，共情家属遭遇，营造出善意的舆论环境。

5 结论 / Conclusion

本研究提出了一种基于 BERTopic 模型对短文本突发事件微博舆情主题演化分析的总体研究框架，对重大突发公共事故“东航 Mu5735 空难事件”进行主题识别，结合生命周期理论和百度指数趋势走向将事件划分为三个阶段，可视化各个阶段的主题及文档分布，展示不同阶段舆情焦点；以划分的三个阶段为时间切片，对相邻阶段的主题采用余弦相似度度量主题间的相似性，以桑基图形式展示在全过程中“东航 Mu5735 空难事件”网络舆情的主题演化路径。

在舆情发生期，网民对事故发生过程、搜救情况、遇难者家属心理疗愈等方面较为关注；在舆情扩散期，舆情主要焦点为失联者皆以遇难的强大悲情评论、事故引发的中美关系外部舆论环境变动、黑匣子的破译进度以及遇难者信息公布与否等这四个方面；在舆情消退期，该阶段的主题聚焦于事故调查以及遇难者缅怀两个方面。借助 BERTopic 模型将各阶段主题可视化，明确每个时期舆情焦点，精准把控突发公共事故舆情发展方向。整个时间的舆情演化是从初发的有关事故基本情况、救援过程关注、失事乘客家属安抚以及事故谣言四类主题演化到扩散期的搜救结果公布、遇难者家属求助以及意愿征求、探究事故真相三个方向，最终演化到消退期事故调查初步报告通报、清明时节悼念遇难者两个方向。

综上，通过 BERTopic 模型以及相似度度量厘清了选取的突发公共事件网络舆情事件的发展过程，挖掘事件焦点主题、分析演化流程，为后续治理此类舆情提供理论分析基础。本研究的局限与展望：本研究当前仅是对于主题进行识别以及分析演化过程，后续可分析该类舆情的情感演化，以及提高数据来源的多样性。

参考文献 / References：

[1] 杜洪涛, 王君泽, 李婕. 基于多案例的突发事件网络舆情演化模式研究[J].情报学报,2017,36(10):1038-1049.

[2] Zhong Zufeng. Internet public opinion evolution in the COVID-19 event and coping strategies[J].Disaster medicine and public health preparedness,2021,15(06):27-33.

[3] Chen Min, Du Wenhu. The predicting public sentiment evolution on public emer

gencies under deep learning and internet of things[J].The Journal of Supercomputing,2023,79(6):6452-6470.

[4] 许露萌. 时空视角下突发事件网络舆情热度演变分析[J].应用数学进展,2022,11(05):3009-3017.

[5] 王晰巍, 王小天, 李玥琪. 重大突发事件网络舆情 UGC 的事理图谱构建研究——以自然灾害 7· 20 河南暴雨为例[J].图书情报工作,2022,66(16):13-23.

[6] 张柳, 王慧, 相甍甍. 基于 LDA 的突发事件应急管理主题热度与演化分析[J].情报科学,2023,41(06):182-191.

[7] 曹树金, 岳文玉. 突发公共卫生事件微博舆情主题挖掘与演化分析[J].信息资源管理学报,2020,10(06):28-37.

[8] 周林兴, 王帅. 事理图谱模型下的重大突发事件网络舆情诱发与缓释机理研究[J].图书情报工作,2023,67(12):58-69.

[9] 杨洋洋. 事件驱动、权威主导与公众诉求：重大突发事件中网络舆情触发机制研究[J].情报资料工作,2023,44(01):33-41.

[10] 杨洋洋, 胡峰. 突发灾难事件中舆情发酵机理与内在逻辑研究——基于融知发酵模型[J].情报杂志,2024,43(03):157-164+25.

[11] 卢国强, 黄微, 刘毅洲. 群体极化视域下突发事件网络舆情极端观点识别研究[J].情报资料工作,2023,44(01):42-51.

[12] 卢国强, 黄微, 杨佩霖. 基于风险耦合的突发事件网络舆情群体极化演化分析[J].现代情报,2023,43(03):96-109.

[13] 贾若男, 王晰巍, 王楠阿雪. 突发事件网络舆情群体极化风险评估研究[J].图书情报工作,2024,68(06):83-92.

[14] Deerwester Scott, Dumais Susan T, Furnas George W. Indexing by Latent Semantic Analysis[J].Journal of the American Society for Information Science,1990,41(6):391-407.

[15] Hofmann Thomas.Probabilistic Latent Semantic Indexing[EB/OL].[2024-04-14].https://dl.acm.org/doi/pdf/10.1145/312624.312649.

[16] Blei David M, Ng Andrew Y, Jordan Michael I. Latent Dirichlet Allocation[J].Journal of Machine Learning Research,2003,3(1):993-1022.

[17] 田世海, 吕德丽. 改进潜在语义分析和支持向量机算法用于突发安全事件舆情预警[J].数据分析与知识发现,2017,1(02):11-18.

[18] 周楠, 杜攀, 靳小龙等. 面向舆情事件的子话题标签生成模型 ET-TAG[J].计算机学报,2018,41(07):1490-1503.

[19] 曾子明, 陈思语. 基于 LDA 与 BERT-BiLSTM-Attention 模型的突发公共卫生

事件网络舆情演化分析[J].情报理论与实践,2023,46(09):158-166.
[20] Yu Dejian, Xiang Bo. Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling[J].Expert Systems with Applications,2023:120114.
[21] Sakshi, Kukreja Vinay. Recent trends in mathematical expressions recognition: An LDA-based analysis[J].Expert Systems with Applications,2023,213:119028.
[22] Angelov Dimo.Top2vec: Distributed representations of topics[EB/OL].[2024-04-10].https://arxiv.org/abs/2008.09470.
[23] Ghasiya Piyush, Okamura Koji. Investigating COVID-19 news across four nations: A topic modeling and sentiment analysis approach[J].Ieee Access,2021,9:36645-36656.
[24] Grootendorst Maarten.BERTopic: Neural topic modeling with a class-based TF-IDF procedure[EB/OL].[2024-02-04].https://arxiv.org/abs/2203.05794.
[25] Wang Zhongyi, Chen Jing, Chen Jiangping et al.Identifying interdisciplinary topics and their evolution based on BERTopic [J/OL]. Scientometrics:1-26 [2024-04-14].https://link.springer.com/article/10.1007/s11192-023-04776-5.
[26] 曹树金, 曹茹烨. 基于研究主题和引文分析的信息资源管理学科发展探究[J].信息资源管理学报,2023,13(2):12-29.
[27] Mcinnes Leland, Healy John, Melville James.Umap: Uniform manifold approximation and projection for dimension reduction[EB/OL].[2024-02-04].https://arxiv.org/abs/1802.03426.

作者贡献声明：

江长斌，黄英辉，王宏宇：提出研究思路，设计研究方案，论文修改意见；
江长斌，徐思思：模型构建，实验设计，论文起草与论文最终版本修订；
徐思思，徐紫琪：采集、清洗和分析数据。

Research on the Evolution of Public Opinion Themes on Microblogs of Emergency Events Based on the BERTopic: A Case Research of the "Eastern Airlines Flight MU5735 Crash"*

Jiang Changbin, Xu Sisi, Huang Yinghui, Xu Ziqi, Wang Hongyu

(School of Management, Wuhan University of Technology, Wuhan Hubei 430070, China)

Abstract: [Purpose/Significance] This research aims to systematically analyze the thematic evolution trends of public sentiment during emergency events, visualizing the focal themes throughout the development process of public sentiment, and providing a practical reference for guiding future online public sentiment. **[Method/Process]** Utilizing the BERTopic model for topic extraction, this research identifies the themes at different stages of sentiment development and measures the similarity between themes using cosine similarity to visualize the thematic evolutionary paths. The case of the "Eastern Airlines Flight Mu5735 crash" on Sina Weibo is examined to research the thematic evolution of public sentiment during an emergency. **[Results/Conclusion]** The empirical results demonstrate that the BERTopic model is effective and offers high visibility in theme identification for sentiment events, accurately capturing the hot topics in each phase of sentiment development and revealing the thematic evolution process during the spread of public sentiment. **[Innovation/Limitation]** In this research, we propose a general framework for analyzing the theme evolution of microblog public opinion on short-text emergencies based on the BERTopic model, and we analyze the theme content evolution of the extracted results and present them visually. The limitation of this study lies in the fact that the data sources selected in this study are only from the microblogging platform, and the diversity of data sources can be improved in the future.

Keywords: Public sentiment in sudden accidents; BERTopic; Theme identification; Thematic evolution

---

* This work is supported by National Social Science Foundation project titled "Research on the Evolutionary Laws of 'Hidden' Political Opinions and Governance Paths under the Perspective of Big Data." (Grant No. 19BSH013)

Author(s): Jiang Changbin，associate professor，master supervisor；Xu Sisi，master's degree candidate；Huang Yinghui，associate professor，doctorate；Xu Ziqi, master's degree candidate；Wang Hongyu，associate researcher，doctorate，corresponding author，E-mail：hongyuwang@whut.edu.cn.